09/08/99
JC674 U.S. PT

A

PTO/SB/05 (2/98) (modified)
Approved for use through 9/30/2000, OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

JC584 U.S. PTO
09/392170
09/08/99

NEW UTILITY PATENT APPLICATION TRANSMITTAL

(only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket Number	3792(PD-595)
First Named Inventor	Monika R. Henzinger
Total Pages in this Submission	74
Express Mail Label No.	EM009695248US

APPLICATION ELEMENTS

1. ☑ Fee Transmittal Form (in duplicate)
   - ☑ Check Enclosed
2. ☑ Specification
   (preferred arrangement set forth below)
   - ■ Descriptive Title of the Invention
   - ■ Cross Reference(s) to Related Case(s)
   - ■ Statement Regarding Fed sponsored R & D
   - ■ Background of the Invention
   - ■ Brief Summary of the Invention
   - ■ Brief Description of the Drawing(s)
   - ■ Detailed Description
   - ■ Claim or Claims
   - ■ Abstract of the Disclosure
3. ☑ Drawing(s) *(when necessary per 35 USC 113)*
4. Oath or Declaration
   a. ☑ New Declaration
      ☑ Executed
   b. ☐ Copy from a prior application (37 CFR 1.63(d))
      (for continuation/divisional with Box 17 completed)
      i. ☐ DELETION OF INVENTOR(S)
         Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. ☐ Incorporation by Reference *(useable if Box 4b is checked).* The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.

ACCOMPANYING APPLICATION PARTS

6. ☑ Assignment & Assignment Recordation Cover Sheet
7. ☐ Certified Copy of Priority Document(s)
   (if foreign priority is claimed)
8. ☐ Information Disclosure Statement & PTO-1449
   ☐ Copies of IDS Citation(s)
9. ☐ Preliminary Amendment
10. Small Entity Statement
    ☐ New Statement enclosed
    ☐ Statement filed in prior application. Status still proper and desired
11. ☑ Return Postcard
12. ☐ ____________________
13. ☐ ____________________
14. ☐ ____________________
15. ☐ ____________________
16. ☐ ____________________

ADDRESS TO:

Assistant Commissioner for Patents
Box Patent Application
Washington, D.C. 20231

17. **If a CONTINUING APPLICATION,** *check appropriate box and supply the requisite information below and in a preliminary amendment:*
☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No: ___/_______
Prior application information: *Examiner:* ____________________ *Group/Art Unit:* ____________

18. CORRESPONDENCE ADDRESS

NAME	Amir H. Raubvogel Fenwick & West LLP				
ADDRESS	Two Palo Alto Square				
CITY	Palo Alto	*STATE*	CA	*ZIP CODE*	94306
COUNTRY	U.S.A.	*TELEPHONE*	(650) 858-7276	*FAX*	(650) 494-1417

Name (Print/Type)	Amir H. Raubvogel	*Registration No. (Attorney/Agent)*	37,070
Signature		*Date*	SEP 8/99

928098
Rev. 09/08/99

RANKING SEARCH ENGINE RESULTS

Inventors:
Monika R. Henzinger
Michael D. Mitzenmacher

Prepared by:

Amir H. Raubvogel
Reg. No. 37,070
Fenwick & West LLP
Two Palo Alto Square
Palo Alto, CA 94306

09392170.090899

RANKING SEARCH ENGINE RESULTS

Inventors:
Monika R. Henzinger
Michael D. Mitzenmacher

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates generally to search engines, and more particularly to a system and method of evaluating and ranking search engines and their results.

09392170.090899

2. Description of Background Art

With the ever-growing size and popularity of the World Wide Web has come an increasingly difficult challenge: providing users with high-quality mechanisms for searching and navigating an enormous and diverse quantity of information. Users attempting to locate information on the Web often begin by running a search on one of several freely-available search engines, such as those found at www.yahoo.com, www.infoseek.com, and the like. Such search engines generally perform some form of keyword search on web documents, and return a list of "hits" representing pages or websites having information relevant to the keyword.

Often, the number of hits returned is very large, and the user is faced with the burdensome task of trying to determine which, if any, of the hits may lead to useful information. Some search engines attempt to rank the hits in order to provide some guidance as to which are more likely to be useful. Such ranking may be based, for example, on the relative prominence of the keyword within the web page, or the number of occurrences of the keyword within the web page. However, it has been found that such ranking techniques are often unreliable, as they do not accurately reflect the relative quality of a particular web page or website.

The relative quality of a web page has been found to be an effective predictor of whether the page will be relevant or useful to a search. Since the World Wide Web is so diverse, with virtually anyone being able to publish pages at will, there is a wide range of quality of pages on the Web. Some pages may be published by large commercial entities with journalistic standards and fact-checking or by academic institutions with scrupulous review procedures, while others may be published by individuals with no quality control, and with no inclination or capability to verify the information being posted. In addition, many web pages employ attention-getting strategies specifically designed to manipulate the page's relative rank in conventional search engines. Since such techniques may be employed by any web page at will, conventional search engines have difficulty assessing relative quality without being given extraneous information regarding the publisher of particular pages and websites.

Quality of a website, while necessarily a subjective term, can however be measured. Page et al. [1], "The PageRank Citation Ranking: Bringing Order to the Web", January 1998, describes a "PageRank" method for measuring the relative importance (or quality) of web pages in order to provide a ranking system based on an objective criterion. In essence, PageRank is a recursive technique which ranks a page based on the sum of the ranks of the pages that link to it. Thus, a page that is linked to by a large number of pages

09903747.030503

tends to be ranked relatively highly, particularly if the linking pages are themselves of high rank. As a precursor to developing PageRank measurements, Page et al. [1] performs a random walk through the Web by following successive links on pages.

However, the PageRank technique suffers from a number of disadvantages. Pages that are part of a large commercial site often contain massive amounts of internal links, to and from other pages within the same site. Such a situation may unduly skew the PageRank results in favor of such pages. Results so ranked may provide the user with a large number of hits from one monolithic source, rather than a diverse array of useful search results. In addition, implementation of Page et al. [1]'s technique involves an initial mapping of the entire document space being indexed, potentially the entire World Wide Web, a substantially daunting and time-consuming task. If the entire document space is not indexed, the PageRank measure may be an inaccurate approximation based on the sub-graph of pages actually indexed.

In addition, users are often faced with a decision as to which of several distinct web search engines to use for a particular search. Various search engines and their associated indexes are themselves of varying degrees of quality, depending on how likely they are to return a result that will be of use to the user. Thus, an overall assessment of the quality of a search engine index as compared with other search engine indexes may offer guidance to a user as to which to use for a particular search.

Traditionally, search engine indexes have been compared with one another based on the size, or number of pages, they contain or index. Such a measure may be of some use, particularly in the context of advertising for a search engine, as size is sometimes considered to be an indicator of retrieval performance for the end user. See, for example, K. Bharat and A. Broder, "A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines", in Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, April 1998, pp. 379-88. However, size of the search engine index is at best a crude indicator of performance, as it fails to take into account the relative quality of the pages that are retrieved by the search engine, which has been found to be of greater importance than the number of pages retrieved.

What is needed is a system and method for ranking search engine indexes and search results, which avoids the above-referenced deficiencies and facilitates retrieval of a diverse collection of high-quality documents. What is further needed is a ranking system and method which does not require mapping out of the entire document space prior to operation. What is further needed is a ranking system and method which avoids the above-referenced problems in comparing pages from a large site containing many internal links with pages from smaller sites. What is further needed is a ranking system and method which measure search engine index quality in an objective manner that considers relative quality of retrieved pages.

09392170 .090899

Summary of the Invention

In accordance with the present invention, there is provided a system and method of measuring and ranking search engine results based on relative quality. The present invention can be used to generate a ranked order of results for a particular search, as well as to perform a comparison of overall quality of a number of search engine indexes.

The present invention employs a two-level random walk in order to generate an improved measure of page quality. In traversing the document space, the present invention treats all pages within a particular grouping (such as a website) as belonging to one node. Selection of the next destination in the random walk is determined first at the node level, and then a particular page within the node is selected. By traversing the document space in this manner, the present invention generates a measurement of quality that is more likely to be based on the number of outside back-links rather than to be skewed by an excessive number of back-links originating within the same website. Thus, documents belonging to large commercial websites having many internal links are not given an unfair advantage in the page ranking.

Search engine index quality can be measured by determining what percentage of documents encountered on the random walk are indexed by the search engine. Document quality can be measured by determining how many times a document is encountered during the random walk; in other words, the more time the random walk spends at a particular document, the higher the relative quality of that document.

The present invention offers other advantages as well. Selected nodes can be treated distinctly from other nodes, depending on some characterization of their relative importance. Thus, a particular node might be excluded from the quality measurement for some reason, or another node might be given greater weight.

In addition, the present invention is able to start measuring the quality of pages without necessarily mapping the entire document space. By employing a random walk, the present invention can determine an approximation of page rank measurement using data for visited pages. Thus, the requirement for advance mapping of the document space is avoided, and searches and page rankings can begin more quickly.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flowchart of a random walk method of sampling pages according to one embodiment of the present invention.

Fig. 2 is a detailed flowchart of a random walk method of sampling pages.

Fig. 3 is an example of a hyperlinked document set.

Fig. 4 is an example of a hyperlinked document set containing hosts of varying sizes.

Fig. 5 is a flowchart showing a method of generating a search engine index quality metric from the output of a random walk.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

For illustrative purposes, the following description presents the invention in the context of web pages and websites that form part of the World Wide Web. However, it will be apparent to one skilled in the art that the present invention can be applied to any set of documents or files residing within a document space or other collection of data. Accordingly, the present invention should not be considered to be limited to a web-based implementation. In addition, the words "page" and "document" are used interchangeably in the context of this invention, to denote any distinct file, entity, or item containing data.

The present invention generates a measure of the quality of a search engine result, both in terms of an individual result for comparison with other results in connection with a particular query, and in terms of the overall quality of a search engine index in comparison with other search engine indexes. Thus, the present invention can be applied, for example, to rank the results of a particular search, as well as to rank the relative quality of several search engine indexes.

For broad queries, a measure of the quality of search engine results can be of significant value. Conventionally, users are often presented with a large number of results (or "hits") for such queries, and are at a loss as to which results to explore first. By providing a measurement of search result quality measurement, the present invention attempts to determine which hits are most likely to be relevant to the user, so as to increase the effectiveness and efficiency of searches.

In one embodiment, the present invention employs a page quality measurement known as the PageRank ranking, as described in S. Brin et al., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 107-17, April 1998. PageRank develops a measurement of the quality of the page based on the number of other pages that link to that page. In another embodiment, the present invention employs an improved version of the PageRank measurement, as described below.

09392170 . 090899

In the World Wide Web, and in other hyperlinked document sets, most pages contain links to other pages. If page A links to page C, then page C is said to be a "back-link" of page A. Thus, the number of back-links of a page, also known as the "InDegree" of the page, is a measure of the number of other pages that point to that page. Generally, pages having a large number of back-links, i.e. a high "InDegree", are considered more important or of higher quality than other pages.

Referring now to Fig. 3, there is shown an example of a hyperlinked document set 300 containing five documents 301-305 illustrating the concepts of back-links and "InDegree". Document 301 contains links pointing to documents 304 and 305, so that document 301 is considered to be a back-link of documents 304 and 305. Similarly, document 302 points to documents 301 and 304, document 303 points to documents 304, document 304 points to documents 302, 303, and 305, and document 305 points to document 303. The InDegree of each document can be determined by counting the number of back-links it contains; thus, documents 301, 302, and 305 have InDegree of 1, while documents 303 and 304 have InDegree of 3.

Furthermore, as described in Brin et al., PageRank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page. A formal definition of the improved PageRank measure as employed in one embodiment of the present invention will be provided below. Intuitively, PageRank approximates the behavior of a "random surfer" who begins at a random web page and continues to click on links in the page, occasionally starting on another random web page. A probability known as a "damping factor" d is defined, specifying the likelihood that the random surfer will request a random page instead of following a link.

Generally, then, a page can be given a high PageRank if many other pages point to it , or if there are some pages that point to it and themselves have a high PageRank.

The present invention extends and improves the PageRank concepts in several ways, as will be described below.

Random Walks

In one embodiment, the present invention derives a measurement of page quality by performing a random walk. If X = {s_1, s_2, ... , s_n} is a set of states, a random walk on X corresponds to a sequence of states, one for each step of the walk. At each step, the walk switches from its current state to a new state or remains at the current state. Random walks are usually Markovian, which signifies that the transition at each step is independent of the previous steps and depends only on the current state.

One embodiment of the present invention utilizes a Markovian random walk on the document set (such as the web), where each page in the document set represents a possible state. For a set of hyperlinked documents, a natural way to move between states is to follow a hyperlink from one page to another.

The equilibrium distribution of the walk is defined as, for each state, the fraction of the steps the random walk would spend in the state if the random walk continued for an infinite amount of time. In most well-behaved walks, the probabilities given by the equilibrium distribution are very closely approximated by the probabilities that one finds a random walk in a given state at some point far, but finitely far, in the future.

Page Quality Measurement

The present invention employs a definition of quality of a search engine index as follows. If each page p of the document set is given a weight w(p), with the weights being scaled so that the sum of all weights is 1, the quality of a search engine index S can be defined as:

$$w(S) = \sum_{p \in S} w(p) \qquad \text{(Eq. 1)}$$

Regardless of the choice of w, according to the above definition the quality of a search engine index is to some extent related to its size. In particular, if the pages indexed by a search engine index S_1 are a subset of the pages indexed by a search engine index S_2, then S_2 will have at least as large a quality score as S_1 by the above criterion. Thus, a second metric, the average page quality of a search engine index, may be employed, defined as:

$$A(S) = w(S) / \ |S| \qquad \text{(Eq. 2)}$$

where $|S|$ is the number of pages indexed by search engine index S.

The average page quality provides an indication of how well a search engine index selects pages to index. However, large search engine indexes are at a disadvantage, since the more pages an index contains, the more difficult it will be to keep the average page quality high.

Average page quality also provides a measurement of relative quality of search results within a particular search engine index, and thus may be used for ranking results returned by a search engine, as will be seen below.

In one embodiment, the present invention utilizes an improved version of the PageRank measure for page quality. As described in Brin et al., the PageRank measure is a quality metric that takes into account not only the number of pages that reference a page, but also the PageRank of the referencing pages as well. This recursive definition provides for a measurement that is in accord with the intuitive concept that links from a high-quality page should be given more weight than links from a low-quality page.

A formal definition of PageRank may be expressed as follows:

$$R(p) = d / T + (1-d)\sum_{i=1}^{k} R(p_i) / C(p_i) \qquad \text{(Eq. 3)}$$

where:

T is the total number of pages in the document set;

d is a damping factor such that 0 < d < 1, with a typical value between, for example, 0.1 and 0.15, though any value might be used;

pages $p_1, \dots, p_k$ link to page p;

R(p) is the PageRank of p; and

C(p) is the number of links out of p.

R(p) can be scaled so that the sum of all R(p) is 1, in which case R(p) can be thought of as a probability distribution over pages and hence a weight function.

As discussed above, PageRank (and the improved version described herein) may be interpreted in terms of the behavior of a "random surfer" who follows links and periodically (depending on the damping factor) selects a random page. The equilibrium probability that such a surfer is at page p is given as R(p). Thus, pages with high rank are more likely to be visited than pages with low rank.

<u>Search Engine Index Quality</u>

In one embodiment, the present invention develops a measurement of search engine index quality by independently selecting pages $p_1, p_2, p_3, \dots,$ p_n in the document set and testing whether each selected page is indexed by the search engine index S. Thus, if the sequence of pages $p_1, p_2, p_3, \dots, p_n$ is the sample sequence, and if $I[p_i \in S]$ is 1 if page p_i is indexed by S, and 0 if not, then an estimate for search engine index quality is given as:

$$\overline{w}(S) = \frac{1}{n}\sum_{i=1}^{n} I[p_i \in S] \qquad \text{(Eq. 4)}$$

Thus, the quality of the search engine index is approximated by the fraction of pages in the sample sequences that is indexed by S. Furthermore, the expectation of each $I[p_i \in S]$ is given by w(S), as follows:

$$E(I[p_i \in S]) = \sum_{p \in S} \Pr(p_i = p) = \sum_{p \in S} w(p) = w(S) \qquad \text{(Eq. 5)}$$

Thus, $\overline{w}(S)$ is the average of several independent binary random variables, each taking the value 1 with probability w(S), which implies that:

$$E(w(S)) = \frac{1}{n}\sum_{i=1}^{n} E(I[p_i \in S]) = w(S) \qquad \text{(Eq. 6)}$$

Thus, the present invention estimates the quality of a search engine index, as well as its results, by selecting pages according to w, and testing whether each selected page is indexed by the search engine index.

In one embodiment, the present invention tests whether a page is indexed by a search engine index as follows. Using a list of words that appear in documents and an approximate measure of their frequency, the invention finds the k rarest words that appear in each document, where k is any number (such as, for example, 9). The search engine is then queried using a conjunction of these k rarest words, and the results are checked to determine whether they include the page. See, for example, Bharat et al.

Referring now to Fig. 1, there is shown a flowchart of a method of sampling pages according to one embodiment of the present invention.

The walk begins with an initial host 106 and random selection 102 of a page within the host. At each step in the random walk, the present invention decides 103 randomly (based on the damping factor) whether to follow a link on the current page or to select a random new page. If following a link, the invention selects 104 a link on the current page and follows it 105 (i.e. retrieves a page corresponding to the link). If selecting a random new page, the invention selects 101 a host uniformly at random from the set of hosts encountered on the walk so far, and selects 102 a page chosen uniformly at random from the set of pages discovered on that host thus far. If, however, a page with no outgoing links is encountered, the page and its host are not recorded, so that the walk is not restarted at a dead end. The loop of Fig. 1 may be repeated until all pages have been traversed, or more likely until some predetermined condition is reached.

The two-level (host, then page) random walk method of Fig. 1 has been found to increase the spread of the walk in comparison with prior art methods, reducing the bias in favor of hosts having large numbers of interconnected pages.

09392170.090999

Referring now to Fig. 4, there is shown an example of a hyperlinked document set 400 containing hosts 401-406 of varying sizes, each host containing one or more documents. Host 401, for example, contains a relatively large number of interconnected documents 410-416, while host 403 contains just two documents 422 and 423. According to prior art methods, a document such as 414, having an InDegree of 6, would be ranked approximately equal to document 422, also having an InDegree of 6 (subject to adjustment based on the InDegrees of the referring documents). The present invention would take into account the fact that document 414 belongs to a large intra-connected host 401, and that the back-links of document 414 come from documents within the same host 401, while the back-links of document 422 come from documents from various hosts. Thus, the relative quality of document 422 is likely to be higher. The two-level random walk method reduces the bias in favor of documents in large hosts such as 401, by reducing the amount of time spent traversing links within a single host and thereby increasing the spread of the walk.

In one embodiment, the present invention keeps track of all visited pages (and their associated hosts) for the purpose of performing a random jump to a previously-visited page. This information may be stored, for example, in random-access memory (RAM) or on secondary storage such as a disk. In an alternative embodiment, a limited number of pages is recorded, such as for example the most recently visited 100,000 pages. In yet another embodiment, only a subset of visited pages are recorded, using a probabilistic sampling method. Such alternative techniques may serve to reduce the storage burden associated with recording all visited pages.

It has been found that any bias resulting from selection of the initial host and page within that host is substantially reduced or eliminated after a sufficiently large number of steps in the walk have been completed. In one embodiment, the first steps in the walk are discarded, so as to reduce such a bias even further. Alternatively, the damping factor can be decreased for early steps in the walk, so as to increase the likelihood that links will be followed rather than attempting to randomly select among relatively few hosts.

One embodiment of the present invention performs random walks using Mercator, an extensible, multi-threaded web crawler written in the Java programming language. In one embodiment, a number of random walks can be conducted in parallel, each walk running in a separate thread of control. When a walk randomly jumps to a page instead of following a link, it can choose a host uniformly at random from all hosts seen by any thread thus far, and then choose a page on that host uniformly from all pages on that host seen by any thread so afar.

In one embodiment, a "host" is defined as a domain containing a set of pages, such as for example www.yahoo.com. However, depending on the nature of the document set, "host" may be defined as any collective group or set of documents.

Referring now to Fig. 2, there is shown a detailed flowchart of the random walk method of sampling pages, as followed by each thread in parallel in one embodiment of the present invention. The following variables are shared by all threads:

HostSet, the set of host names discovered so far;

UrlSet(h), the set of Uniform Resource Locators (URLs) or other document identifiers, discovered so far that belong to host h; and

Samples, a list of URLs representing the sample sequence.

The system starts 200 by assigning initial values to HostSet, UrlSet, and Samples. For example, HostSet may be set to a popular website such as www.yahoo.com; UrlSet(www.yahoo.com) may be set to {www.yahoo.com}; UrlSet(h) may be set to {} for all other hosts h; and Samples may be set to [].

The system selects 201 a host h uniformly at random from HostSet. Next, it selects 202 a URL u uniformly at random from UrlSet(h), the URL set associated with the selected host. The system then downloads 203 the page p referred to by u, using conventional downloading means.

In 204, the system determines whether page p contains at least one link. If so, steps 205 through 209 are performed. The system assigns 205 h to be equal to the host component of URL u (i.e., that portion of URL u that identifies a particular host). If, in step 206, h is in HostSet, the system, in step 207, adds h to HostSet. If, in step 208, u is in UrlSet(h), the system, in step 209, adds u to UrlSet(h). If in step 204, the system determined that page p did not contain any links, the system proceeds to step 210.

In 210, with probability c, the system adds u to Samples. In 211, the system determines whether to attempt to follow a link on page p (by proceeding to 212) or, with probability d, to return to step 201 to select a new host at random.

In 212, the system assigns U to represent the set of URLs (links) contained in page p. If in 213, U is empty, the system returns to step 201 to select a new host. If in 213, U is not empty, the system proceeds to step 214.

In 214, the system chooses and removes a URL u uniformly at random from U. In 215, the system attempts to download page p referred to by u. If redirects are encountered, they are followed. In one embodiment, the present invention limits the number of consecutive HTTP redirects to, for example, five, in order to avoid redirect cycles.

In one embodiment, the system favors links that are external to the current host h, so as to increase the likelihood of visiting a large number of different hosts rather than remaining within the same host.

If in 216, the attempted download was unsuccessful, the system returns to step 213. If the download was successful, the system determines 217 whether the downloaded page is an HTML page. In one embodiment, the present invention only uses pages that are HTML pages, and ignores pages that do not have a content type of "text/html" in the HTTP response header. If the page is not HTML, the system returns to step 213.

If the downloaded page is HTML, the system returns to step 204 to begin the cycle again at the next step.

The steps of Fig. 2 can be repeated any number of times, until it is determined that sufficient iterations have been completed or until some system limitation is reached. Based on the results of the random walk, relative quality of individual pages can be determined so that search results can be ranked accordingly. In essence, the more often a page is visited during the random walk, the higher its quality ranking. This implies that pages that are referenced by high-quality pages are also given higher quality rankings. Furthermore, as described previously, relative quality of search engine index quality can be determined by measuring the number of high-quality pages referenced by the search engine index.

It has been found that the two-level random walk yields improved results by avoiding biases in favor of large intraconnected sites. In addition, page quality measurement can occur without requiring indexing of the entire document set in advance, as a ranking can be based on the pages visited so far in the random walk at any given time. Furthermore, individual hosts or other sets of pages can be singled out for exclusion from the random walk, or special weight, or other special treatment, as desired.

Given the random walk described above, a rank measure can be generated for each page to be indexed. In one embodiment, the rank measure is developed from the two-level random walk in a similar manner as described by Page et al. [1] and for conventional random walks. Further details of the PageRank measure are found, for example, in Page et al. [1]; and Page et al. [2], "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in To

Appear: Proceedings of the Seventh International Web Conference (WWW 98), 1998.

As discussed above, the relative quality of a search engine index can be estimated from the output generated by the random walk, by determining what fraction of pages encountered in the random walk are indexed by the search engine. Referring now to Fig. 5, there is shown a flowchart of a technique for generating a search engine index quality metric, given the output of the random walk described above. The system begins by initializing i=0 and N=0. It then selects 501 a URL from Samples (see above). If in 502, the selected URL is indexed by the search engine index, the system increments i 503. N is incremented 504 regardless of whether the selected URL is indexed. If more URLs exist 505, the system returns to 501. Once all URLs in Samples have been processed, the system outputs i/N 506, which represents the fraction of URLs from Samples that were indexed, and therefore provides an indication of the quality of the search engine index. This value can then be used to compare search engine indexes with one another.

The output of the random walk can also be used to determine a quality metric for each page encountered on the walk. The number of times a particular page is encountered is an indication of the page's quality. This value can be normalized as follows:

$$\text{Quality(page)} = (\text{\# of times page appears}) / (\text{Total \# of steps in walk}) \quad \text{(Eq. 7)}$$

Thus, the quality is described in terms of the fraction of all steps in the walk that are spent at a particular page.

From the above description, it will be apparent that the invention disclosed herein provides a novel and advantageous system and method of evaluating and ranking search engine indexes and their results. The foregoing discussion discloses and describes merely exemplary methods and embodiments of the present invention. As will be understood by those familiar with the art, the invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. Accordingly, the disclosure of the present invention is intended to be illustrative, but not limiting, of the scope of the invention, which is set forth in the following claims.

What is Claimed is:

1. A computer-implemented method for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, the method comprising:

a) selecting a host;

b) selecting at random a document associated with the host;

c) retrieving the selected document;

d) selecting at random a link in the retrieved document;

e) retrieving a document referenced by the selected link; and

f) repeating d) and e) until a predetermined condition is met.

2. The method of claim 1, further comprising, prior to d):

c.1) responsive to a random event:

c.1.1) selecting at random a host from among the previously selected hosts; and

c.1.2) repeating b) through f);

and wherein f) comprises repeating c.1) through e) until a predetermined condition is met

3. The method of claim 1, further comprising, prior to d):

c.1) generating a random number;

c.2) determining whether the random number falls within a predetermined range; and

c.3) responsive to the random number falling within the predetermined range:

c.1.1) selecting at random a host from among the previously selected hosts; and

c.1.2) repeating b) through f).

4. The method of claim 1, wherein the document set is the World Wide Web, and wherein each document is a web page.

5. The method of claim 4, wherein each host corresponds to a domain.

6. The method of claim 1, further comprising, concurrently with a) through f), performing a second two-level random walk through the hypertext-linked document set.

7. A computer-implemented method for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, the method comprising:

a) initializing a host set;

b) initializing a document set for each host in the host set;

c) selecting at random a host from the host set;

d) selecting at random a document from the document set of the selected host;

e) adding the selected host to the host set;

f) adding the selected document to the document set of the selected host;

g) responsive to the selected document containing at least one link:

g.1) selecting at random a link from the selected document;

g.2) selecting a document corresponding to the selected link;

g.3) selecting a host corresponding to the selected document;

g.4) repeating e) through h) until a predetermined condition is met; and

h) responsive to the selected document not containing at least one link, repeating c) through h) until a predetermined condition is met.

8. The method of claim 7, wherein:

e) is performed responsive to the selected host not being in the host set; and

f) is performed responsive to the selected document not being in the document set of the selected host.

9. The method of claim 7, wherein g) further comprises, prior to g.1):

g.0) responsive to a random event, repeating c) through h) until a predetermined condition is met;

and wherein g.1) through g.4) are performed responsive to non-occurrence of the random event of g.0).

10. The method of claim 7, further comprising, prior to g.1):

g.0.1) generating a random number;

g.0.2) determining whether the random number falls within a predetermined range; and

g.0.3) responsive to the random number falling within the predetermined range, repeating c) through h) until a predetermined condition is met;

and wherein g.1) through g.4) are performed responsive to the random number not falling within a predetermined range.

11. The method of claim 7, wherein the hypertext-linked document set is the World Wide Web, and wherein each document is a web page.

12. The method of claim 11, wherein each host corresponds to a domain.

13. A computer-implemented method for measuring relative quality of a search engine index, comprising:

a) performing a two-level random walk among documents within a document set;

b) for each document encountered in the random walk, determining whether the document is indexed by the search engine index; and

c) aggregating the results of b).

14. The method of claim 13, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, and wherein a) comprises:

a.1) selecting a host;

a.2) selecting at random a document associated with the host;

a.3) retrieving the selected document;

a.4) selecting at random a link in the retrieved document;

a.5) retrieving a document referenced by the selected link; and

a.6) repeating a.4) and a.5) until a predetermined condition is met.

15. The method of claim 14, further comprising, prior to a.4):

a.3.1) responsive to a random event:

a.3.1.1) selecting at random a host from among the previously selected hosts; and

a.3.1.2) repeating a.2) through a.6).

16. The method of claim 13, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, and wherein a) comprises:

a.1) initializing a host set;

a.2) initializing a document set for each host in the host set;

a.3) selecting at random a host from the host set;

a.4) selecting at random a document from the document set of the selected host;

a.5) adding the selected host to the host set;

a.6) adding the selected document to the document set of the selected host;

a.7) responsive to the selected document containing at least one link:

- a.7.1) selecting at random a link from the selected document;
- a.7.2) selecting a document corresponding to the selected link;
- a.7.3) selecting a host corresponding to the selected document;
- a.7.4) repeating a.5) through a.8) until a predetermined condition is met; and

a.8) responsive to the selected document not containing at least one link, repeating a.3) through a.8) until a predetermined condition is met.

17. The method of claim 16, wherein:

a.5) is performed responsive to the selected host not being in the host set; and

a.6) is performed responsive to the selected document not being in the document set of the selected host.

18. The method of claim 13, wherein each document contains a plurality of words, and wherein b) comprises, for each document encountered in the random walk:

b.1) selecting at least one word from the document;

b.2) performing a query on the search engine index based on the selected at least one word, to obtain search results; and

b.3) determining whether the document is included in the obtained search results.

19. The method of claim 18, wherein b.1) comprises selecting at least one word based on rarity.

09392170 .090899

20. A computer-implemented method for measuring relative quality of a document in a document set, comprising:

a) performing a two-level random walk among documents within a document set; and

b) determining a quality metric responsive to the number of times the document is encountered in the random walk.

21. A computer-implemented method for measuring relative quality of a document in a document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, the method comprising:

a) performing a two-level random walk among documents within a document set; and

b) determining a quality metric responsive to the number of documents that link to the document.

22. The method of claim 21, wherein b) comprises determining a quality metric responsive to the number of documents that link to the document, and responsive to the quality metric of the linking documents.

23. The method of claim 21, wherein b) comprises determining a value for:

$$R(p) = d / T + (1-d)\sum_{i=1}^{k} R(p_i) / C(p_i)$$

where:

T is the total number of documents in the document set;

d is a damping factor such that $0 < d < 1$;

documents $p_1, \dots, p_k$ each contain at least one link to document p; and

C(p) is the number of links out of p.

24. The method of claim 21, wherein each document is associated with a host, and wherein a) comprises:

a.1) selecting a host;

a.2) selecting at random a document associated with the host;

a.3) retrieving the selected document;

a.4) responsive to a random event:

a.4.1) selecting at random a host from among the previously selected hosts; and

a.4.2) repeating a.2) through a.7);

a.5) selecting at random a link in the retrieved document;

a.6) retrieving a document referenced by the selected link; and

a.7) repeating a.4) to a.6) until a predetermined condition is met.

25. The method of claim 21, wherein each document is associated with a host, and wherein a) comprises:

a.1) initializing a host set;

a.2) initializing a document set for each host in the host set;

a.3) selecting at random a host from the host set;

a.4) responsive to a random event:

a.4.1) selecting at random a host from among the previously selected hosts; and

a.4.2) repeating a.2) through a.7).

a.5) selecting at random a document from the document set of the selected host;

a.6) adding the selected host to the host set;

a.7) adding the selected document to the document set of the selected host;

a.8) responsive to the selected document containing at least one link:

a.8.1) selecting at random a link from the selected document;

a.8.2) selecting a document corresponding to the selected link;

a.8.3) selecting a host corresponding to the selected document; and a.8.4) repeating a.6) through a.9) until a predetermined condition is met; and

a.9) responsive to the selected document not containing at least one link, repeating a.3) through a.9) until a predetermined condition is met.

26. The method of claim 21, further comprising:

c) determining a quality metric for at least one additional document; and

d) ranking the quality metric of the first document with respect to the quality metrics of the additional documents.

27. A computer-implemented method for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, the method comprising:

a) selecting a host;

b) selecting at random a document associated with the host;

c) retrieving the selected document;

d) responsive to a random event:

d.1) selecting at random a host from among the previously selected hosts; and

d.2) repeating b) through e) until a predetermined condition is met

e) responsive to the random event not occurring:

e.1) selecting at random a link in the retrieved document;

e.2) retrieving a document referenced by the selected link; and

e.3) repeating d) and e) until a predetermined condition is met.

28. A computer-implemented method for measuring relative quality of a document in a document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, the method comprising:

a) performing a two-level random walk among documents within a document set, the two-level random walk comprising:

a.1) initializing a host set;

a.2) initializing a document set for each host in the host set;

a.3) selecting at random a host from the host set;

a.4) responsive to a random event:

a.4.1) selecting at random a host from among the previously selected hosts; and

a.4.2) repeating a.2) through a.7).

a.5) selecting at random a document from the document set of the selected host;

a.6) adding the selected host to the host set;

09392170 .090899

a.7) adding the selected document to the document set of the selected host;

a.8) responsive to the selected document containing at least one link:

a.8.1) selecting at random a link from the selected document;

a.8.2) selecting a document corresponding to the selected link;

a.8.3) selecting a host corresponding to the selected document;

a.8.4) repeating a.6) through a.9) until a predetermined condition is met; and

a.9) responsive to the selected document not containing at least one link, repeating a.3) through a.9) until a predetermined condition is met;

b) determining a quality metric responsive to the number of documents that link to the document;

c) determining a quality metric for at least one additional document; and

d) ranking the quality metric of the first document with respect to the quality metrics of the additional documents.

29. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of

links to other documents, each document being associated with a host, the computer program product comprising:

a) computer-readable program code devices configured to cause a computer to select a host;

b) computer-readable program code devices configured to cause a computer to select at random a document associated with the host;

c) computer-readable program code devices configured to cause a computer to retrieve the selected document;

d) computer-readable program code devices configured to cause a computer to select at random a link in the retrieved document;

e) computer-readable program code devices configured to cause a computer to retrieve a document referenced by the selected link; and

f) computer-readable program code devices configured to cause a computer to repeat the operations of d) and e) until a predetermined condition is met.

09092170.090899

30. The computer program product of claim 29, further comprising computer-readable program code devices configured to cause a computer to, prior to selecting at random a link in the retrieved document:

c.1) responsive to a random event:

select at random a host from among the previously selected hosts; and

repeat the operations of b) through f);

and wherein the computer-readable program code devices configured to cause a computer to repeat the operations of d) and e) until a predetermined condition is met comprise computer-readable program code devices configured to cause a computer to repeat the operations of c.1) through e) until a predetermined condition is met.

31. The computer program product of claim 29, further comprising:

computer-readable program code devices configured to cause a computer to generate a random number;

computer-readable program code devices configured to cause a computer to determine whether the random number falls within a predetermined range; and

computer-readable program code devices configured to cause a computer to, responsive to the random number falling within the predetermined range:

select at random a host from among the previously selected hosts; and

repeat the operations of b) through f).

32. The computer program product of claim 29, wherein the document set is the World Wide Web, and wherein each document is a web page.

33. The computer program product of claim 32, wherein each host corresponds to a domain.

34. The computer program product of claim 29, further comprising computer-readable program code devices configured to cause a computer to,

concurrently with the operations of a) through f), perform a second two-level random walk through the hypertext-linked document set.

35. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, the computer program product comprising:

a) computer-readable program code devices configured to cause a computer to initialize a host set;

b) computer-readable program code devices configured to cause a computer to initialize a document set for each host in the host set;

c) computer-readable program code devices configured to cause a computer to select at random a host from the host set;

d) computer-readable program code devices configured to cause a computer to select at random a document from the document set of the selected host;

e) computer-readable program code devices configured to cause a computer to add the selected host to the host set;

f) computer-readable program code devices configured to cause a computer to add the selected document to the document set of the selected host;

g) computer-readable program code devices configured to cause a computer to, responsive to the selected document containing at least one link:

g.1) select at random a link from the selected document;

g.2) select a document corresponding to the selected link;

g.3) select a host corresponding to the selected document; and

g.4) repeat the operations of e) through h) until a predetermined condition is met; and

h) computer-readable program code devices configured to cause a computer to, responsive to the selected document not containing at least one link, repeat the operations of c) through h) until a predetermined condition is met.

09580560.041730

36. The computer program product of claim 35, wherein:

the computer-readable program code devices configured to cause a computer to add the selected host to the host set operate responsive to the selected host not being in the host set; and

the computer-readable program code devices configured to cause a computer to add the selected document to the document set of the selected host operate responsive to the selected document not being in the document set of the selected host.

37. The computer program product of claim 35, wherein computer-readable program code devices g) further comprise computer-readable program code devices configured to cause a computer to, prior to g.1):

g.0) responsive to a random event, repeat the operations of c) through h) until a predetermined condition is met;

and wherein computer-readable program code devices g) are configured to cause a computer to perform g.1) through g.4) responsive to non-occurrence of the random event of g.0).

09392170.090899

38. The computer program product of claim 35, wherein computer-readable program code devices g) further comprise computer-readable program code devices configured to cause a computer to, prior to g.1):

g.0.1) generate a random number;

g.0.2) determine whether the random number falls within a predetermined range; and

g.0.3) responsive to the random number falling within the predetermined range, repeat the operations of c) through h) until a predetermined condition is met;

and wherein computer-readable program code devices g) are configured to cause a computer to perform g.1) through g.4) responsive to the random number not falling within a predetermined range.

39. The computer program product of claim 35, wherein the hypertext-linked document set is the World Wide Web, and wherein each document is a web page.

40. The computer program product of claim 39, wherein each host corresponds to a domain.

41. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for measuring relative quality of a search engine index, the computer program product comprising:

a) computer-readable program code devices configured to cause a computer to perform a two-level random walk among documents within a document set;

b) computer-readable program code devices configured to cause a computer to, for each document encountered in the random walk, determine whether the document is indexed by the search engine index; and

c) computer-readable program code devices configured to cause a computer to aggregate the results of the operations of b).

42. The computer program product of claim 41, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, and wherein the computer-readable program code devices configured to cause a computer to perform a two-level random walk comprise:

a.1) computer-readable program code devices configured to cause a computer to select a host;

a.2) computer-readable program code devices configured to cause a computer to select at random a document associated with the host;

a.3) computer-readable program code devices configured to cause a computer to retrieve the selected document;

a.4) computer-readable program code devices configured to cause a computer to select at random a link in the retrieved document;

a.5) computer-readable program code devices configured to cause a computer to retrieve a document referenced by the selected link; and

a.6) computer-readable program code devices configured to cause a computer to repeat the operations of a.4) and a.5) until a predetermined condition is met.

43. The computer program product of claim 42, further comprising computer-readable program code devices configured to cause a computer to, prior to selecting at random a link in the retrieved document:

a.3.1) responsive to a random event:

select at random a host from among the previously selected hosts; and

repeat the operations of a.2) through a.6).

44. The computer program product of claim 41, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, and wherein the computer-readable program code devices configured to cause a computer to perform a two-level random walk comprise:

a.1) computer-readable program code devices configured to cause a computer to initialize a host set;

a.2) computer-readable program code devices configured to cause a computer to initialize a document set for each host in the host set;

a.3) computer-readable program code devices configured to cause a computer to select at random a host from the host set;

a.4) computer-readable program code devices configured to cause a computer to select at random a document from the document set of the selected host;

a.5) computer-readable program code devices configured to cause a computer to add the selected host to the host set;

a.6) computer-readable program code devices configured to cause a computer to add the selected document to the document set of the selected host;

a.7) computer-readable program code devices configured to cause a computer to, responsive to the selected document containing at least one link:

a.7.1) select at random a link from the selected document;

a.7.2) select a document corresponding to the selected link;

a.7.3) select a host corresponding to the selected document;

a.7.4) repeat the operations of a.5) through a.8) until a predetermined condition is met; and

a.8) computer-readable program code devices configured to cause a computer to, responsive to the selected document not containing at least one link, repeat the operations of a.3) through a.8) until a predetermined condition is met.

45. The computer program product of claim 44, wherein:

the computer-readable program code devices configured to cause a computer to add the selected host to the host set are configured to cause a computer to add the selected host responsive to the selected host not being in the host set; and

the computer-readable program code devices configured to cause a computer to add the selected document to the document set of the selected host are configured to cause a computer to add the selected document responsive to the selected document not being in the document set of the selected host.

46. The computer program product of claim 41, wherein each document contains a plurality of words, and wherein the computer-readable program code devices configured to cause a computer to, determine whether the document is indexed by the search engine index comprise computer-readable

program code devices configured to, for each document encountered in the random walk:

b.1) select at least one word from the document;

b.2) perform a query on the search engine index based on the selected at least one word, to obtain search results; and

b.3) determine whether the document is included in the obtained search results.

47. The computer program product of claim 46, wherein the computer-readable program code devices configured to select at least one word from the document comprise computer-readable program code devices configured to select at least one word based on rarity.

48. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for measuring relative quality of a document in a document set, the computer program product comprising:

computer-readable program code devices configured to cause a computer to perform a two-level random walk among documents within a document set; and

computer-readable program code devices configured to cause a computer to determine a quality metric responsive to the number of times the document is encountered in the random walk.

49. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for measuring relative quality of a document in a document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, the computer program product comprising:

computer-readable program code devices configured to cause a computer to perform a two-level random walk among documents within a document set; and

computer-readable program code devices configured to cause a computer to determine a quality metric responsive to the number of documents that link to the document.

50. The computer program product of claim 49, wherein the computer-readable program code devices configured to cause a computer to determine a quality metric comprise computer-readable program code devices configured to cause a computer to determine a quality metric responsive to the number of documents that link to the document, and responsive to the quality metric of the linking documents.

51. The computer program product of claim 49, wherein the computer-readable program code devices configured to cause a computer to determine a quality metric comprise computer-readable program code devices configured to cause a computer to determine a value for:

$$R(p) = d / T + (1-d)\sum_{i=1}^{k} R(p_i) / C(p_i)$$

where:

09390432.021199

T is the total number of documents in the document set;

d is a damping factor such that 0 < d < 1;

documents $p_1, \dots, p_k$ each contain at least one link to document p; and

C(p) is the number of links out of p.

52. The computer program product of claim 49, wherein each document is associated with a host, and wherein the computer-readable program code devices configured to cause a computer to perform a two-level random walk comprise:

a.1) computer-readable program code devices configured to cause a computer to select a host;

a.2) computer-readable program code devices configured to cause a computer to select at random a document associated with the host;

a.3) computer-readable program code devices configured to cause a computer to retrieve the selected document;

a.4) computer-readable program code devices configured to cause a computer to, responsive to a random event:

a.4.1) select at random a host from among the previously selected hosts; and

a.4.2) repeat the operations of a.2) through a.7);

a.5) computer-readable program code devices configured to cause a computer to select at random a link in the retrieved document;

a.6) computer-readable program code devices configured to cause a computer to retrieve a document referenced by the selected link; and

a.7) computer-readable program code devices configured to cause a computer to repeat the operations of a.4) to a.6) until a predetermined condition is met.

53. The computer program product of claim 49, wherein each document is associated with a host, and wherein and wherein the computer-readable program code devices configured to cause a computer to perform a two-level random walk comprise:

a.1) computer-readable program code devices configured to cause a computer to initialize a host set;

a.2) computer-readable program code devices configured to cause a computer to initialize a document set for each host in the host set;

a.3) computer-readable program code devices configured to cause a computer to select at random a host from the host set;

a.4) computer-readable program code devices configured to cause a computer to, responsive to a random event:

a.4.1) select at random a host from among the previously selected hosts; and

a.4.2) repeat the operations of a.2) through a.7).

a.5) computer-readable program code devices configured to cause a computer to select at random a document from the document set of the selected host;

a.6) computer-readable program code devices configured to cause a computer to add the selected host to the host set;

a.7) computer-readable program code devices configured to cause a computer to add the selected document to the document set of the selected host;

a.8) computer-readable program code devices configured to cause a computer to, responsive to the selected document containing at least one link:

a.8.1) select at random a link from the selected document;

a.8.2) select a document corresponding to the selected link;

a.8.3) select a host corresponding to the selected document; and

a.8.4) repeat the operations of a.6) through a.9) until a predetermined condition is met; and

a.9) responsive to the selected document not containing at least one link, repeating the operations of a.3) through a.9) until a predetermined condition is met.

54. The computer program product of claim 49, further comprising:

c) computer-readable program code devices configured to cause a computer to determine a quality metric for at least one additional document; and

d) computer-readable program code devices configured to cause a computer to rank the quality metric of the first document

09690560.041702

with respect to the quality metrics of the additional documents.

55. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, the computer program product comprising:

a) computer-readable program code devices configured to cause a computer to select a host;

b) computer-readable program code devices configured to cause a computer to select at random a document associated with the host;

c) computer-readable program code devices configured to cause a computer to retrieve the selected document;

d) computer-readable program code devices configured to cause a computer to, responsive to a random event:

d.1) select at random a host from among the previously selected hosts; and

d.2) repeat the operations of b) through e) until a predetermined condition is met

e) computer-readable program code devices configured to cause a computer to, responsive to the random event not occurring:

e.1) select at random a link in the retrieved document;

e.2) retrieve a document referenced by the selected link; and

e.3) repeat the operations of d) and e) until a predetermined condition is met.

56. A computer program product comprising a computer-usable medium having computer-readable code embodied therein for measuring relative quality of a document in a document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, the computer program product comprising:

a) computer-readable program code devices configured to cause a computer to perform a two-level random walk among documents within a document set, the computer-readable program code devices comprising:

a.1) computer-readable program code devices configured to cause a computer to initialize a host set;

a.2) computer-readable program code devices configured to cause a computer to initialize a document set for each host in the host set;

a.3) computer-readable program code devices configured to cause a computer to select at random a host from the host set;

a.4) computer-readable program code devices configured to cause a computer to, responsive to a random event:

a.4.1) select at random a host from among the previously selected hosts; and

09392170.090899

a.4.2) repeat the operations of a.2) through a.7).

a.5) computer-readable program code devices configured to cause a computer to select at random a document from the document set of the selected host;

a.6) computer-readable program code devices configured to cause a computer to add the selected host to the host set;

a.7) computer-readable program code devices configured to cause a computer to add the selected document to the document set of the selected host;

a.8) computer-readable program code devices configured to cause a computer to, responsive to the selected document containing at least one link:

a.8.1) select at random a link from the selected document;

a.8.2) select a document corresponding to the selected link;

a.8.3) select a host corresponding to the selected document;

a.8.4) repeat the operations of a.6) through a.9) until a predetermined condition is met; and

a.9) computer-readable program code devices configured to cause a computer to, responsive to the selected document not containing at least one link, repeat the operations of a.3) through a.9) until a predetermined condition is met;

b) computer-readable program code devices configured to cause a computer to determine a quality metric responsive to the number of documents that link to the document;

c) computer-readable program code devices configured to cause a computer to determine a quality metric for at least one additional document; and

d) computer-readable program code devices configured to cause a computer to rank the quality metric of the first document with respect to the quality metrics of the additional documents.

57. A system for randomly walking through a hypertext-linked document set comprising a plurality of documents, wherein at least a subset of the documents contain a plurality of links to other documents, each document being associated with a host, the system comprising:

a) a host selector;

b) a random document selector, coupled to the host selector, for selecting at random a document associated with the host;

c) a document retriever, coupled to the random document selector, for retrieving the selected document; and

d) a link selector, coupled to the document retriever, for selecting at random a link in the retrieved document;

wherein the document retriever retrieves a document referenced by the selected link;

and wherein the link selector repeatedly selects at random a link and the document retriever repeatedly retrieves a document referenced by the selected link, until a predetermined condition is met.

58. A system for measuring relative quality of a search engine index, comprising:

a random walker, for performing a two-level random walk among documents within a document set;

a determination module, coupled to the random walker, for, for each document encountered in the random walk, determining whether the document is indexed by the search engine index; and

a results aggregation module, coupled to the determination module, for aggregating the results of the determination module.

59. A system for measuring relative quality of a document in a document set, comprising:

a random walker, for performing a two-level random walk among documents within a document set; and

a determination module, coupled to the random walker, for determining a quality metric responsive to the number of times the document is encountered in the random walk.

RANKING SEARCH ENGINE RESULTS

ABSTRACT OF THE DISCLOSURE

A method, system, and computer program product for determining relative quality of search engine indexes and search results include performing a two-level random walk through a hypertext-linked document set. Search engine index quality is measured based on the number of encountered documents that are indexed by the search engine index. Search result quality is measured based on the number and quality of documents that link to the result document.

09392170.090899

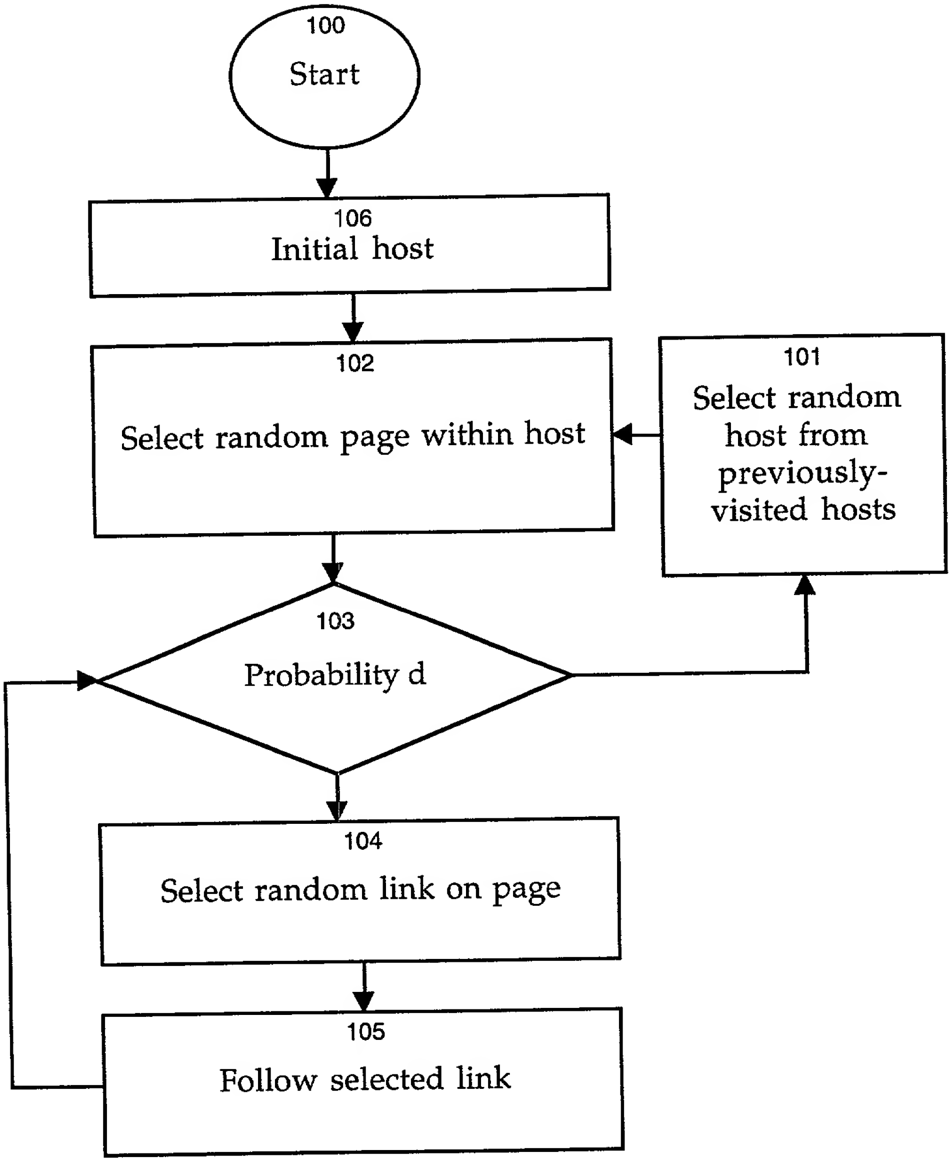
100
Start
106
Initial host
102
Select random page within host
101
Select random host from previously-visited hosts
103
Probability d
104
Select random link on page
105
Follow selected link

FIG. 1

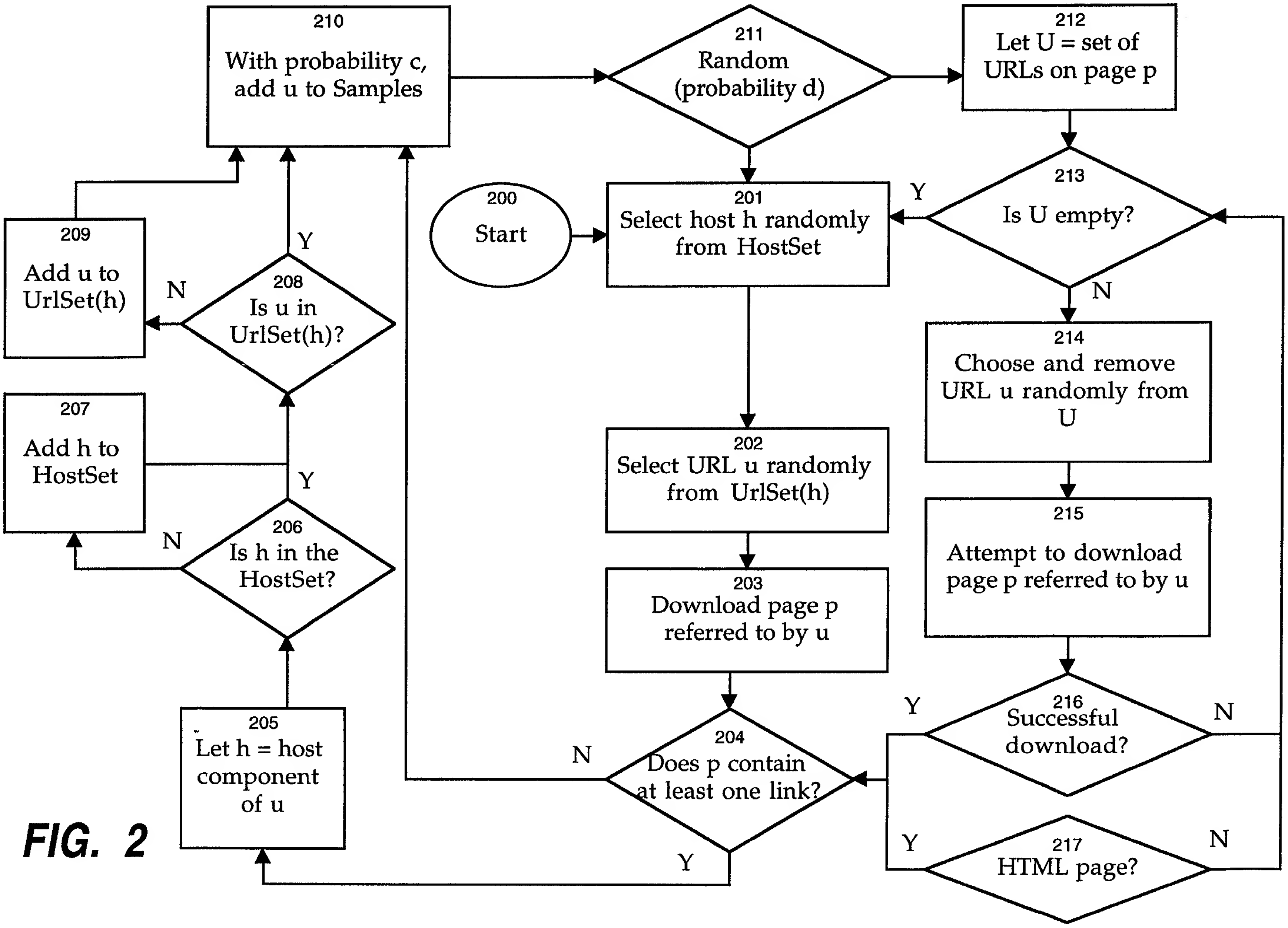
210
With probability c, add u to Samples
211
Random (probability d)
212
Let U = set of URLs on page p
213
Is U empty?
Y
N
200
Start
201
Select host h randomly from HostSet
209
Add u to UrlSet(h)
N
208
Is u in UrlSet(h)?
Y
214
Choose and remove URL u randomly from U
207
Add h to HostSet
Y
202
Select URL u randomly from UrlSet(h)
N
206
Is h in the HostSet?
215
Attempt to download page p referred to by u
203
Download page p referred to by u
205
Let h = host component of u
N
204
Does p contain at least one link?
Y
216
Successful download?
Y
N
217
HTML page?
Y
N

FIG. 2

09392170.090899

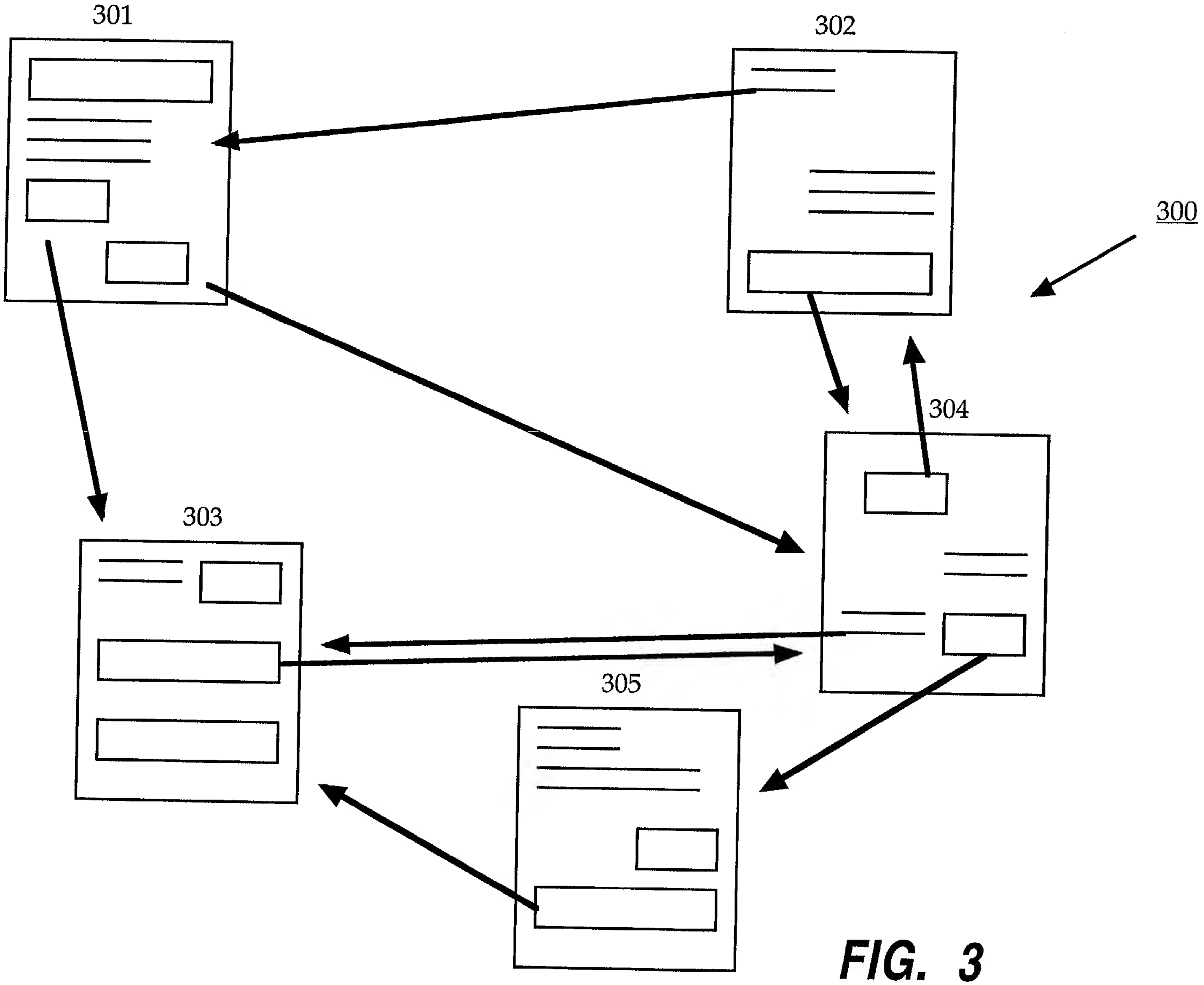

FIG. 3

09392170.090699

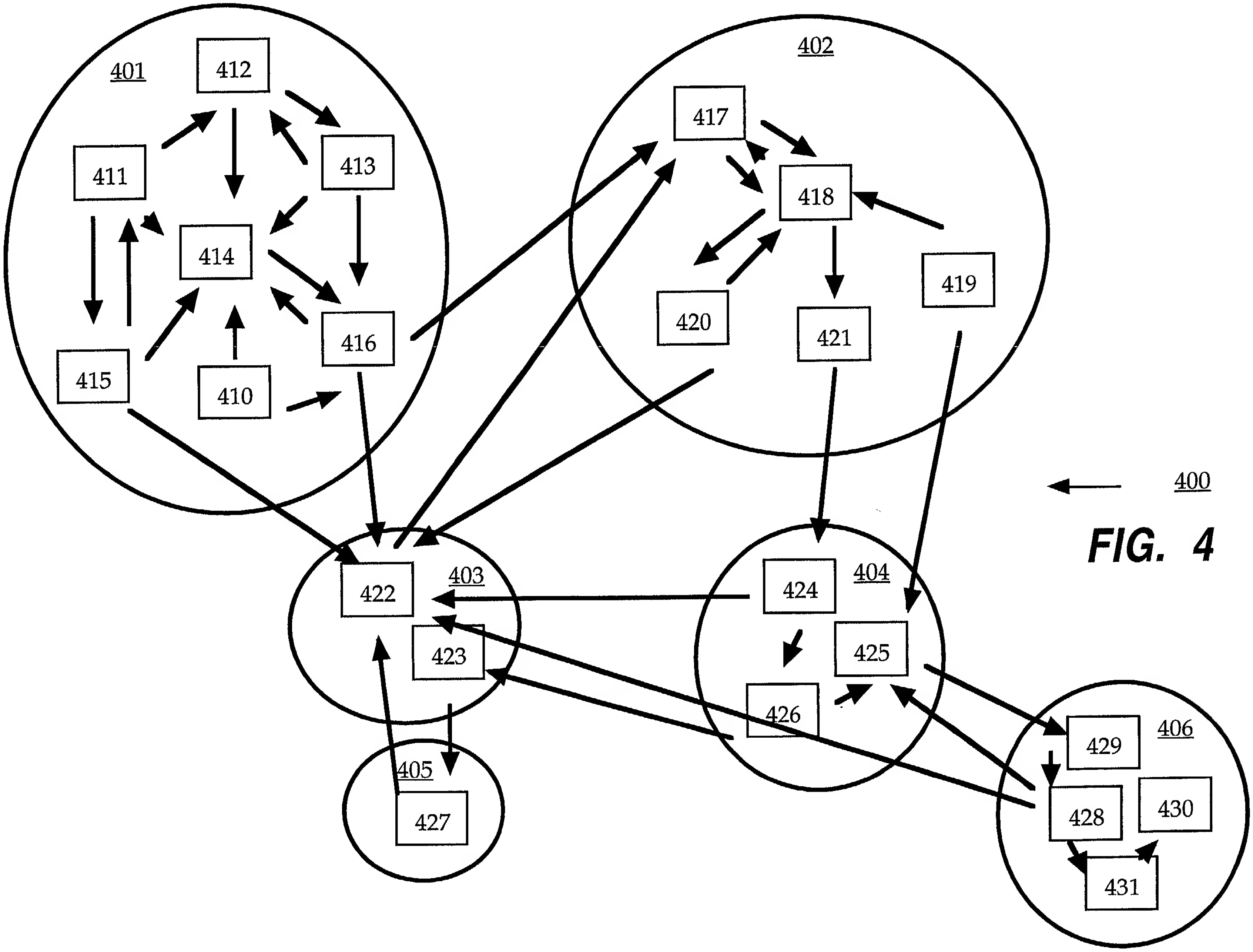

FIG. 4

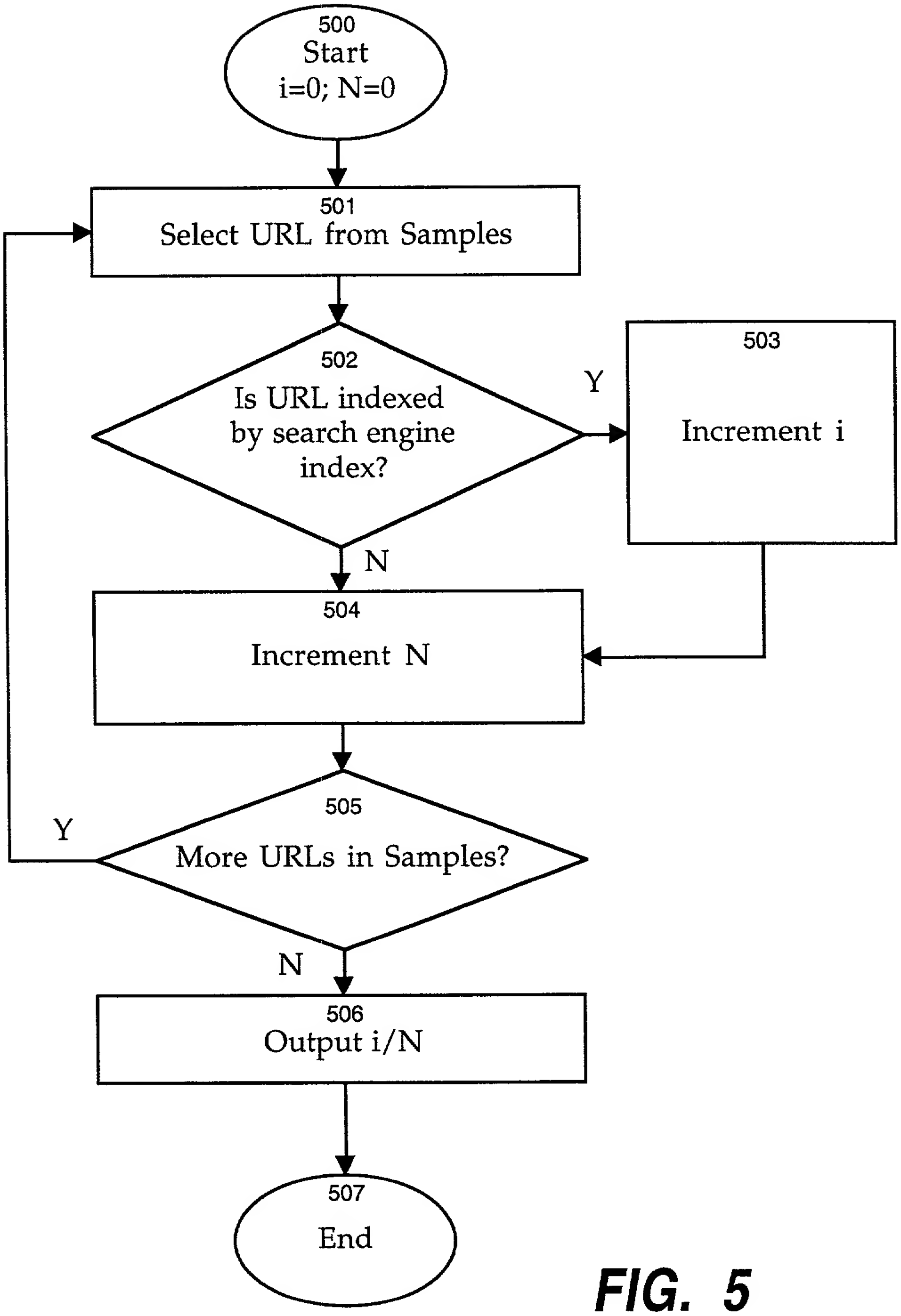
500
Start
i=0; N=0
501
Select URL from Samples
502
Is URL indexed by search engine index?
Y
503
Increment i
N
504
Increment N
505
More URLs in Samples?
Y
N
506
Output i/N
507
End

FIG. 5

PTO/SB/01 (6–95) (modified)
Approved for use through 10/31/96 OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

0010/PTO Rev. 6/95 U.S. Department of Commerce Patent and Trademark Office	Attorney Docket Number	3792 (PD-595)
DECLARATION FOR UTILITY OR DESIGN PATENT APPLICATION	First Named Inventor	**Monika R. Henzinger**
	COMPLETE IF KNOWN	
	Application Number	**not yet known**
	Filing Date	**not yet known**
	Group Art Unit	**not yet known**
[X] Declaration Submitted with Initial Filing OR [ ] Declaration Submitted after Initial Filing	Examiner Name	**not yet known**

As a below named inventor, I hereby declare that:

My residence, post office address, and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

RANKING SEARCH ENGINE RESULTS

the specification of which (*Title of the Invention*)

[x] is attached hereto

OR

[ ] was filed on (MM/DD/YYYY) [____________] as United States Application Number or PCT International Application Number [__________] and was amended on (MM/DD/YYYY) [__________] (if applicable).

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment specifically referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37 Code of Federal Regulations. § 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code § 119 (a)-(d) or § 385(b) of any foreign application(s) for patent or inventor's certificate, or § 365 (a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or of any PCT international application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application Number(s)	Country	Foreign Filing Date (MM/DD/YYYY)	Priority Not Claimed	Certified Copy Attached? YES	NO
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]

[ ] Additional foreign application numbers are listed on a supplemental priority sheet attached hereto:

I hereby claim the benefit under Title 35, United States Code § 119(e) of any United States provisional application(s) listed below.

Application Number(s)	Filing Date (MM/DD/YYYY)	[ ] Additional provisional application numbers are listed on a supplemental sheet attached hereto.

DECLARATION	Page 2

I hereby claim the benefit under Title 35, United States Code § 120 of any United States application(s), or § 365(c) of any PCT international application designating the United States of America, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT international application in the manner provided by the first paragraph of Title 35, United States Code § 112, I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations § 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application.

U.S. Parent Application Number	PCT Parent Number	Parent Filing Date (MM/DD/YYYY)	Parent Patent Number (*if applicable*)

[ ] Additional U.S. or PCT international application numbers are listed on a supplemental priority sheet attached hereto.

As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith:

Name	Registration Number	Name	Registration Number
Lawrence W. Granatelli	**32,228**	**John T. McNelis**	**37,186**
Tina M. Lessani	**41,150**	**Sarah T. Harris**	**35,891**
Richard P. Lange	**27,296**	**Irene Kosturakis**	**33,724**
Joseph Arrambide	**39,589**	**Keith Lutsch**	**31,851**
Barry Blount	**35,069**	**Theodore S. Park**	**26,971**
Amir H. Raubvogel	**37,070**	**Laura A. Majerus**	**33,417**

[ ] Additional attorney(s) and/or agent(s) named on a supplemental sheet attached hereto.

Please direct all correspondence to:

Amir H. Raubvogel
Fenwick & West LLP
Two Palo Alto Square
Palo Alto, CA 94306
U.S.A.

Telephone	**(650) 858-7276**	Fax	**(650) 494-1417**

09536093.041400

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Name of Sole or First Inventor: [ ] A petition has been filed for this unsigned inventor

Given Name	MONIKA	Middle Initial	R.	Family Name	HENZINGER	Suffix e.g. Jr.	

Inventor's Signature		Date	

Residence: City	**Menlo Park**	State	**CA**	Country	**U.S.A.**	Citizenship	**Germany**

Mailing Address	**80 La Loma Drive**
Mailing Address	

City	**Menlo Park**	State	**CA**	Zip	**94025**	Country	**U.S.A.**

[X] Additional inventors are being named on supplemental sheet(s) attached hereto

DECLARATION	ADDITIONAL INVENTOR(S) Supplemental Sheet

Name of Additional Joint Inventor, if any:	[ ] A petition has been filed for this unsigned inventor								
Given Name	MICHAEL	Middle Initial	D.	Family Name	MITZENMACHER	Suffix e.g. Jr.			
Inventor's Signature	Michael D. Mitzenmacher	Date	September 1, 1999						
Residence: City	Belmont	State	MA	Country	U.S.A.	Citizenship	U.S.A.		
Mailing Address	105 Hammond Road # 1								
Mailing Address									
City	Belmont	State	MA	Zip	02478	Country	U.S.A.		

Name of Additional Joint Inventor, if any:	[ ] A petition has been filed for this unsigned inventor						
Given Name		Middle Initial		Family Name		Suffix e.g. Jr.	
Inventor's Signature		Date					
Residence: City		State		Country		Citizenship	
Mailing Address							
Mailing Address							
City		State		Zip		Country	

Name of Additional Joint Inventor, if any:	[ ] A petition has been filed for this unsigned inventor						
Given Name		Middle Initial		Family Name		Suffix e.g. Jr.	
Inventor's Signature		Date					
Residence: City		State		Country		Citizenship	
Mailing Address							
Mailing Address							
City		State		Zip		Country	

Name of Additional Joint Inventor, if any:	[ ] A petition has been filed for this unsigned inventor						
Given Name		Middle Initial		Family Name		Suffix e.g. Jr.	
Inventor's Signature		Date					
Residence: City		State		Country		Citizenship	
Mailing Address							
Mailing Address							
City		State		Zip		Country	

[ ] Additional inventors are being named on supplemental sheet(s) attached hereto

PTO/SB/01 (6–95) (modified)
Approved for use through 10/31/96 OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

0010/PTO Rev. 6/95	U.S. Department of Commerce Patent and Trademark Office	Attorney Docket Number	3792 (PD-595)
		First Named Inventor	Monika R. Henzinger
		COMPLETE IF KNOWN	
		Application Number	not yet known
		Filing Date	not yet known
		Group Art Unit	not yet known
		Examiner Name	not yet known

DECLARATION FOR UTILITY OR DESIGN PATENT APPLICATION

[X] Declaration Submitted with Initial Filing OR [ ] Declaration Submitted after Initial Filing

As a below named inventor, I hereby declare that:

My residence, post office address, and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

RANKING SEARCH ENGINE RESULTS

the specification of which (*Title of the Invention*)

[x] is attached hereto

OR

[ ] was filed on (MM/DD/YYYY) [____________] as United States Application Number or PCT International Application Number [__________] and was amended on (MM/DD/YYYY) [____________] (if applicable).

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment specifically referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37 Code of Federal Regulations. § 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code § 119 (a)-(d) or § 385(b) of any foreign application(s) for patent or inventor's certificate, or § 365 (a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or of any PCT international application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application Number(s)	Country	Foreign Filing Date (MM/DD/YYYY)	Priority Not Claimed	Certified Copy Attached? YES	NO
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]
			[ ]	[ ]	[ ]

[ ] Additional foreign application numbers are listed on a supplemental priority sheet attached hereto:

I hereby claim the benefit under Title 35, United States Code § 119(e) of any United States provisional application(s) listed below.

Application Number(s)	Filing Date (MM/DD/YYYY)	[ ] Additional provisional application numbers are listed on a supplemental sheet attached hereto.

DECLARATION	Page 2

I hereby claim the benefit under Title 35, United States Code § 120 of any United States application(s), or § 365(c) of any PCT international application designating the United States of America, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT international application in the manner provided by the first paragraph of Title 35, United States Code § 112, I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations § 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application.

U.S. Parent Application Number	PCT Parent Number	Parent Filing Date (MM/DD/YYYY)	Parent Patent Number (*if applicable*)

[ ] Additional U.S. or PCT international application numbers are listed on a supplemental priority sheet attached hereto.

As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith:

Name	Registration Number	Name	Registration Number
Lawrence W. Granatelli	**32,228**	**John T. McNelis**	**37,186**
Tina M. Lessani	**41,150**	**Sarah T. Harris**	**35,891**
Richard P. Lange	**27,296**	**Irene Kosturakis**	**33,724**
Joseph Arrambide	**39,589**	**Keith Lutsch**	**31,851**
Barry Blount	**35,069**	**Theodore S. Park**	**26,971**
Amir H. Raubvogel	**37,070**	**Laura A. Majerus**	**33,417**

[ ] Additional attorney(s) and/or agent(s) named on a supplemental sheet attached hereto.

Please direct all correspondence to:

Amir H. Raubvogel
Fenwick & West LLP
Two Palo Alto Square
Palo Alto, CA 94306
U.S.A.

Telephone	**(650) 858-7276**	Fax	**(650) 494-1417**

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Name of Sole or First Inventor: [ ] A petition has been filed for this unsigned inventor

Given Name	MONIKA	Middle Initial	R.	Family Name	HENZINGER	Suffix e.g. Jr.			
Inventor's Signature	Monika Henzinger			Date	8/31/99				
Residence: City	**Menlo Park**	State	**CA**	Country	**U.S.A.**	Citizenship	**Germany**		
Mailing Address	**80 La Loma Drive**								
Mailing Address									
City	**Menlo Park**	State	**CA**	Zip	**94025**	Country	**U.S.A.**		

[X] Additional inventors are being named on supplemental sheet(s) attached hereto

09392170 .090899

DECLARATION	**ADDITIONAL INVENTOR(S)** **Supplemental Sheet**

Name of Additional Joint Inventor, if any:				[ ] A petition has been filed for this unsigned inventor						
Given Name	**MICHAEL**	Middle Initial	**D.**	Family Name	**MITZENMACHER**			Suffix e.g. Jr.		
Inventor's Signature						Date				
Residence: City	**Belmont**	State	**MA**	Country	**U.S.A.**	Citizenship	**U.S.A.**			
Mailing Address	**105 Hammond Road # 1**									
Mailing Address										
City	**Belmont**	State	**MA**	Zip	**02478**	Country	**U.S.A.**			

Name of Additional Joint Inventor, if any:				[ ] A petition has been filed for this unsigned inventor						
Given Name		Middle Initial		Family Name				Suffix e.g. Jr.		
Inventor's Signature						Date				
Residence: City		State		Country		Citizenship				
Mailing Address										
Mailing Address										
City		State		Zip		Country				

Name of Additional Joint Inventor, if any:				[ ] A petition has been filed for this unsigned inventor						
Given Name		Middle Initial		Family Name				Suffix e.g. Jr.		
Inventor's Signature						Date				
Residence: City		State		Country		Citizenship				
Mailing Address										
Mailing Address										
City		State		Zip		Country				

Name of Additional Joint Inventor, if any:				[ ] A petition has been filed for this unsigned inventor						
Given Name		Middle Initial		Family Name				Suffix e.g. Jr.		
Inventor's Signature						Date				
Residence: City		State		Country		Citizenship				
Mailing Address										
Mailing Address										
City		State		Zip		Country				
[ ] Additional inventors are being named on supplemental sheet(s) attached hereto										

09392170 .090899